IGR204 - EuropeDisease

DELIN Jia - DILOU Sully - CALVET Rodolphe
TIZAF Zakaria - BINUANI Nicolas

29 juin 2021

Table des matières

1 Introduction

This document presents the final project we conducted in the IGR204 Data Visualization course.

As seen in this course, we can summarize data visualization as the art of transforming raw data into understandable visual representations for professionals or individuals.

To verify and apply this, we were asked for this project to choose a data set to visualize, to exploit it, to clean it, and even to complete it in order to get useful information. The objective is to understand the users of the visualization, the type of task they are trying to accomplish with the visualization, their business expertise and the type of data they are dealing with.

For this purpose we have chosen a dataset on causes of death in the European Union between 2001 and 2010 from the WHO. It is therefore clear that the users will be health professionals, pharmaceutical companies, and even the ministries of the countries in charge of public health policies.

Our report is divided into five main parts, in which we will present our data, the objective of the target users, the choice of our design, then we will present our visualizations, and finally we will highlight the strengths and limitations of the visualizations.

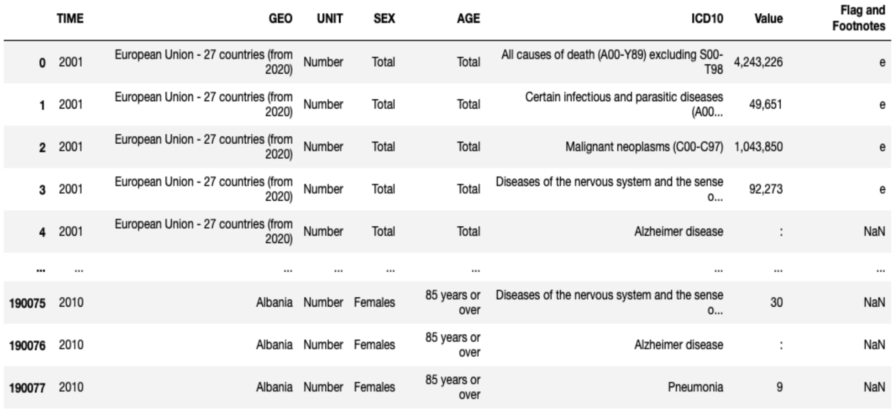
2 Dataset

In this project, we will use mainly two datasets. The first dataset **CauseOfDeath.csv** gives the total number of deaths by disease, gender, age, and year from 2001 to 2010 in Europe. It allows us to have an overview of the distribution of cause of death in European countries and to have a finer study as a function of age, gender etc. This is also the most interesting dataset for our study.

And the second dataset **countries_codes.xlsx** contains country name and its corresponding code for over 200 countries. It aims to visualize each country on the map for Altair. These two datasets provide us the information about the cause of death and the location of country. Now, let's have a closer look at these datasets.

2.1 The Dataset for Cause of Death

Here is an overview of our dataset :



	TIME	GEO	UNIT	SEX	AGE	ICD10	Value	Flag and Footnotes
0	2001	European Union - 27 countries (from 2020)	Number	Total	Total	All causes of death (A00-Y89) excluding S00-T98	4,243,226	e
1	2001	European Union - 27 countries (from 2020)	Number	Total	Total	Certain infectious and parasitic diseases (A00...	49,651	e
2	2001	European Union - 27 countries (from 2020)	Number	Total	Total	Malignant neoplasms (C00-C97)	1,043,850	e
3	2001	European Union - 27 countries (from 2020)	Number	Total	Total	Diseases of the nervous system and the sense o...	92,273	e
4	2001	European Union - 27 countries (from 2020)	Number	Total	Total	Alzheimer disease	:	NaN
...	...	...	...	...	...	...	...	...
190075	2010	Albania	Number	Females	85 years or over	Diseases of the nervous system and the sense o...	30	NaN
190076	2010	Albania	Number	Females	85 years or over	Alzheimer disease	:	NaN
190077	2010	Albania	Number	Females	85 years or over	Pneumonia	9	NaN

Figure 1 – Overview of Dataset

The dataset is organized as a table comprising 190080 rows and 8 columns, for a total of 8 different causes of death (one per row). For our project, 6 features are most useful among all the features and give us different aspect of comprehension : TIME, GEO, SEX, AGE, ICD10, Value. We will discuss these features one by one in the following paragraphs.

TIME : It is a quantitative variable and represent the year of event, it covers from 2001 to 2010, so it takes 10 different values which allows us to study the evolution of cause of death.

GEO : It is a qualitative variable which represents the 27 countries or unions in Europe : 'European Union − 27 countries (from 2020)', 'European Union − 28 countries (2013−2020)', 'Belgium', 'Bulgaria', 'Czechia', 'Denmark', 'Germany (until 1990 former territory of the FRG)', 'Estonia', 'Ireland', 'Greece', 'Spain', 'France', 'France (metropolitan)', 'Croatia', 'Italy', 'Cyprus', 'Latvia', 'Lithuania', 'Luxembourg', 'Hungary', 'Malta', 'Netherlands', 'Austria', 'Poland', 'Portugal', 'Romania', 'Slovenia', 'Slovakia', 'Finland', 'Sweden', 'Iceland', 'Norway', 'Switzerland', 'United Kingdom', 'North Macedonia', 'Albania'. We should mention that this column contains not only the countries but also the European union, so we need to make a distinguish between them when we do the visualization.

SEX : It is nominal(qualitative) variable which takes 3 values : 'Total', 'Males', 'Females'. This column can help us make a comparison between different genders.

AGE : It is another qualitative variable which contains 22 different values : 'Total', 'Less than 1 year', 'From 1 to 4 years','From 5 to 9 years', 'From 10 to 14 years', 'Less than 15 years', 'From 15 to 19 years', 'From 15 to 24 years','From 20 to 24 years', 'From 25 to 29 years', 'From 30 to 34 years', 'From 35 to 39 years','From 40 to 44 years', 'From 45 to 49 years', 'From 50 to 54 years', 'From 55 to 59 years','From 60 to 64 years', 'From 65 to 69 years','From 70 to 74 years', 'From 75 to 79 years','From 80 to 84 years', '85 years or over'. These age ranges allow us to study the rate of death for each cause for each age range.
However, we have modified and removed some of the "AGE" variables for the visualization of the 3rd graph which gives the comparison of the distribution of diseases between men and women at different age groups. The several modifications made are the following :

- We removed the following age groups :
  - 'Less than 1 year'
  - 'From 1 to 4 years'
  - 'Less than 15 years', because we already have the age groups under 15 years
  - 'From 15 to 24 years', because we already have the age groups 'From 15 to 19 years' & 'From 20 to 24 years'

- We have added 'From 0 to 4 years', which is the merging of the age groups 'Less than 1 year' & 'From 1 to 4 years'

- We have renamed '85 years or over' to 'From 85 years to over'

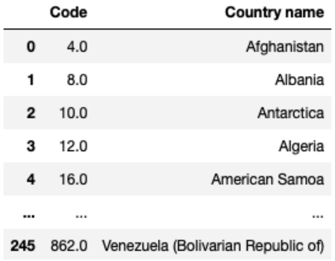
ICD10 : ICD−10 is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. This is the most import qualitative variable in the dataset because it represents the cause of death.

- All causes of death (A00−Y89) excluding S00−T98. S00−T98 : Injury, poisoning and certain other consequences of external causes
- Certain infectious and parasitic diseases (A00−B99)
- Malignant neoplasms (C00−C97)
- Diseases of the nervous system and the sense organs (G00−H95)
- Alzheimer disease
- Pneumonia
- Accidents (V01−X59, Y85, Y86)
- Transport accidents (V01−V99, Y85)

Value : It is a quantitative value which represents the number of death for each line(case).

2.2 Countries codes Dataset

Here is an overview of our dataset :



	Code	Country name
0	4.0	Afghanistan
1	8.0	Albania
2	10.0	Antarctica
3	12.0	Algeria
4	16.0	American Samoa
...	...	...
245	862.0	Venezuela (Bolivarian Republic of)

Figure 2 – Overview of Dataset

The dataset is organized as a table comprising 250 rows and 2 columns, for a total of 250 different countries. The main purpose of using this dataset is to visualize the distribution of number of deaths on a map. This dataset allows to add a column of corresponding country's code in the CauseOfDeath dataset and then to visualize these countries with Altair.

2.3 Separation of Dataset

In order to study different situations, we need to create different data frame so that each visualization can take different frames.

- The first data frame takes only the column with gender 'Total'. Because we want to have an overview of the distribution of the number of deaths first. Then we create a pivot table with the cause of death as index so that we can select a cause to visualize.

- The second data frame allows us to study the evolution of a cause, so this time we choose time and country as our index so that we can see the number of deaths of each country as a function of time.

- Our final data frame is constructed only with gender 'Females' and 'Males' in order to have a comparison between different genders. Moreover, we set age and gender as our index so that we can not only see the difference between genders but also between ages.

3 Target User

We already know that the cause of deaths for people of different genders, ages, and even regions may vary greatly. However, this thought remains fuzzy to public. Moreover, imagine you are working in a pharmacy company or in WHO and you need to decide a strategy for your next steps drug research, or you are a politician, you want to bring benefits to your people by adjusting the health plan so that you can gain more supports for your next election. In both ways, you want to have a closer look at the dataset of European cause of death, to find the evolution of each disease and then make the decision.

The goal of this project is to help doctors, pharmacy companies better understand the evolution of cause of deaths during a decade, especially the causes related to a certain disease, and then to ameliorate or validate some health plans, drug research, and to improve the quality of life for public even further.

Given this user profile and the dataset, we thought someone considering our application might be willing to answer questions along three fundamental axes : discovery/comparison, evolution/ comparison, male/female with age ranges.

3.1 Discovery / Comparison

The question often posed by the doctors or pharmacy companies is that if there is a difference of distribution of some disease in some countries. If yes, what will be the difference between them. Our visualization allows to reply to questions like this :

- Which country has the greatest number of deaths for Alzheimer disease ?
- What is the difference of number of deaths for France and Spain for Transport accidents (V01−V99, Y85) ?
- How many people died by Pneumonia in Norway ?
- Are there any countries that have a distribution of death caused by Malignant neoplasms (C00−C97) like this of Luxembourg ?

3.2 Evolution / Comparison

The second question is about the evolution of each disease in different countries. We all know that the evolution of one disease in a country can attract its people's attention. For example, in USA, unhealthy lifestyle caused a raise of diabetes diseases which attracted not only its people but also other countries' attention. Our visualization allows to reply to questions like this :

- Which country has the most obvious raise of number of deaths for Alzheimer disease ?

- What is the difference of evolution of number of deaths for France and Spain for Transport accidents (V01−V99, Y85) ?
- What is the evolution of Pneumonia in Norway ?
- How many people died in 2003 caused by certain infectious and parasitic diseases (A00−B99) in France ?
- Is there any countries that have an evolution of death caused by Malignant neoplasms (C00−C97) similar to this of Luxembourg ?

3.3 Male / Female with age ranges

Our final question is about the difference of number of deaths for different genders among different age ranges. Nowadays, with the advance of technology, we found that some diseases are gender−selected(genie), males and females have different risk for different diseases. So we would like to elaborate the doctors and researchers for this question.

- Is there a difference of number for different genders for deaths for Alzheimer disease ?
- What is the difference of age distribution for Transport accidents (V01−V99,Y85) ?

4 Chosen design

4.1 Idea that guided in our visualization design

Our goal on this project is to have a tool for health experts, doctors, and politicians about evolution of some most known disease. In fact, we thought about CovidTracker application that is a little like a kind of dashboard to follow disease evolution. So because it's disease repartition between Europe countries, we first think about a chloropleth map that is obvious with this kind of problematic. This is a kind of representation that we want to use :

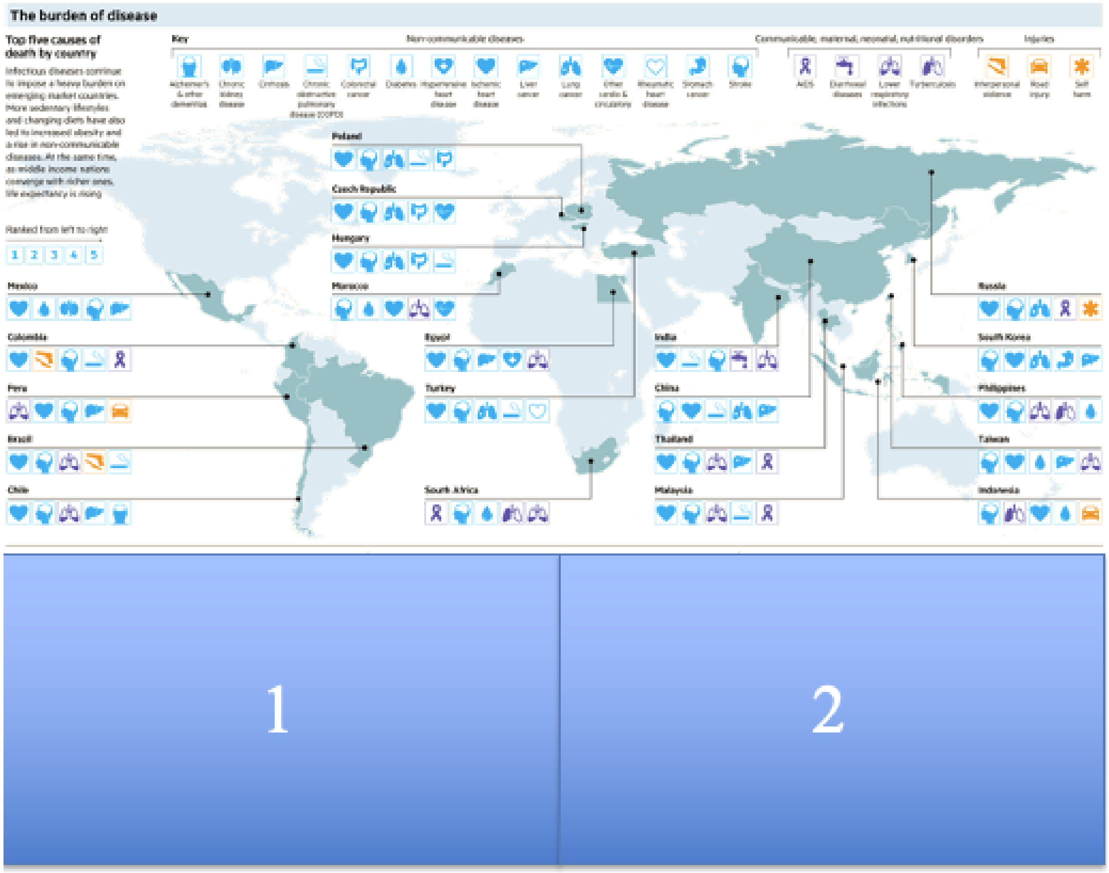


FIGURE 3 – Idea of dashboard for Europe disease

We have a map which we can select different countries and focus on some disease available in our database. Thus to explore more deeper we think about different plots.

4.2 Representations and interactions : map design

In first case we think about a Europe map to represent disease. For example, in our first approach, we suggest a map with disease most represented in each country. See below for a first example of such representation.

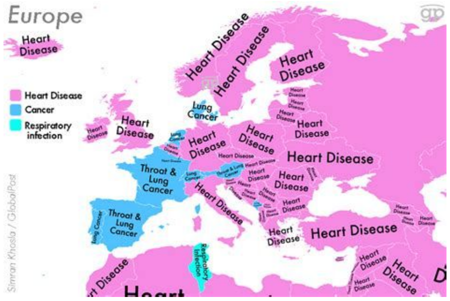


Figure 4 – Idea of map for Europe disease

A problematic in this first idea is that some disease in our dataset are most represented in many countries, therefore maybe two or three diseases will be overexposed and it's not our goal. Remember we want to help politicians (OMS) and medical supplies get more insights, so we more want to focus on one disease and see the repartition between each country. Another idea by adding a plot is representing below.

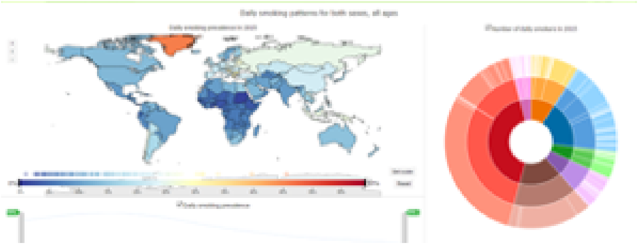
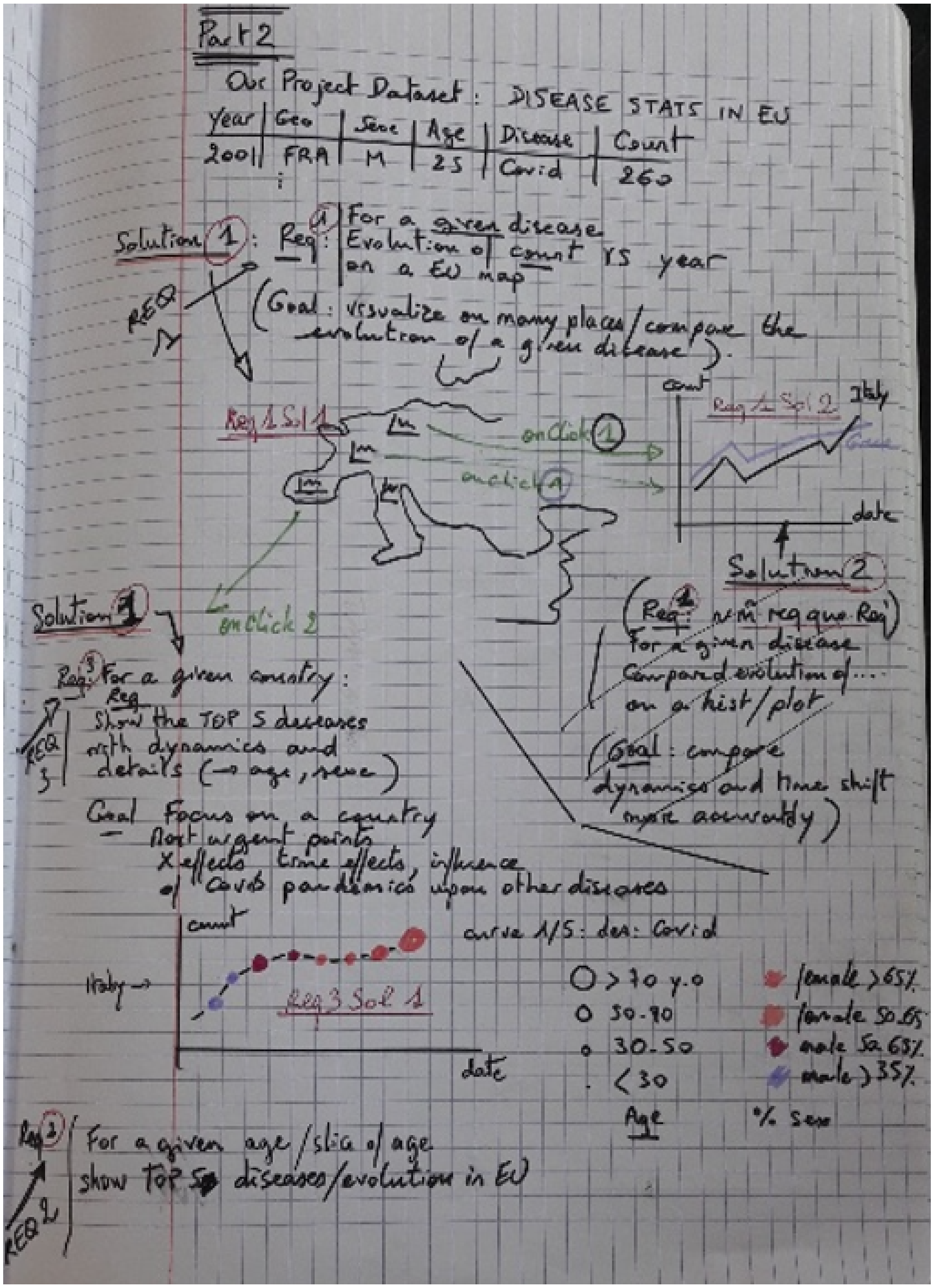


Figure 5 – Idea of map for Europe disease

Now let's start with our own drawn visualization. We begin with our map and add a ListBox.

- Solution1 Request1 : we have for a given disease, the evolution of count VS year on a EU map.
- Solution2 Request2 : we have for a given country, we show top 5 diseases with dynamics details.
- Solution3 Request3 : we have for a given age/slice of age, we show top 5 diseases



Part 2

Our Project Dataset : DISEASE STATS IN EU

Year	Geo	Sexe	Age	Disease	Count
2001	FRA	M	25	Covid	260
	⋮				

Solution 1 : Req 1: For a given disease
Evolution of count VS year
on a EU map

(Goal : visualize on many places / compare the evolution of a given disease).

Solution 2
(Req 1: v.m. req que Req)
For a given disease
Compared evolution of …
on a hist/plot
(Goal : compare dynamics and time shift more accurately)

Solution 1
Req 3: For a given country :
Req
Show the TOP 5 diseases
with dynamics and
details (→ age, sexe)

Goal Focus on a country
– Most urgent points
X effects time effects, influence
of Covid pandemics upon other diseases

curve 1/5 : dea: Covid

Age		% Sexe	
○	> 70 y.o	●	female > 65%
○	50-70	●	female 50-65
∘	30-50	●	male 50-65%
·	< 30	●	male > 35%

Req 2 For a given age / slice of age
show TOP 5 diseases/evolution in EU

Figure 6 – Drawings about our first idea

In a second approach, we think about add animated color legend like representing in following drawing sketchs :

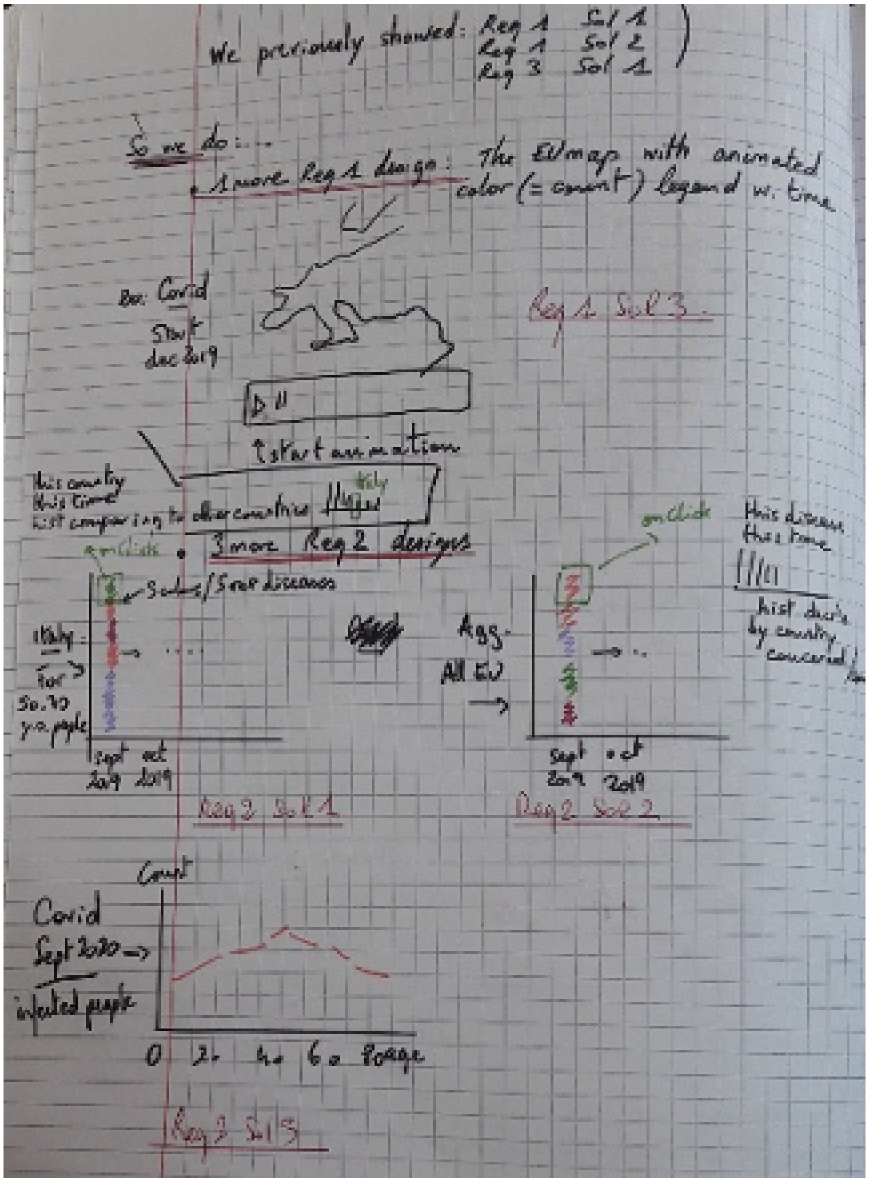


Figure 7 – Drawings about second idea

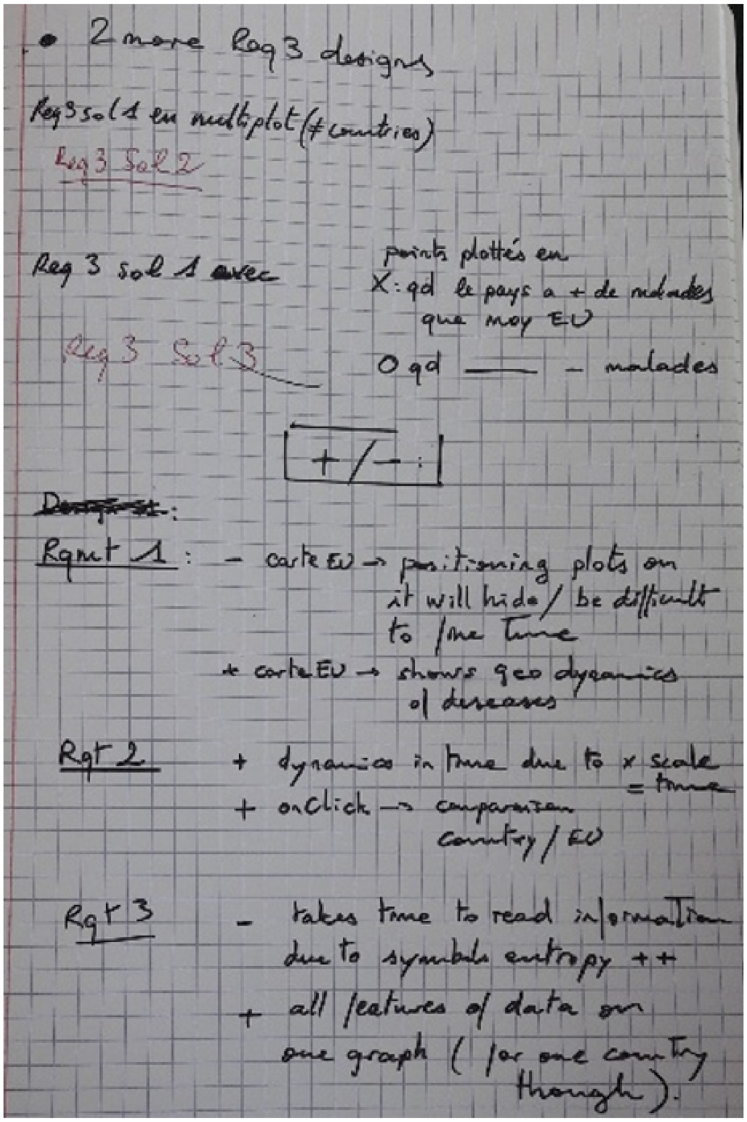
So to conclude on this first approaches, we decided to have a chloropleth map with a selected disease from a ListBox. We add 2 others plots like a dashboard visualization. After some Solution Request sketches we think for examples on the following plots :

- We can also have a barplot for one disease chosen and have a kind of ranking with a barplot / or a rectangular diagram board to focus on which country is more affected (refer to part II.3)
- We can have gender like the example, and we can focus on age repartition depending on disease observed and for each country. (Refer to part II.3)
- We can select one country on a map and have a time serie plot of one disease also selected and for each year point size depending on age patient.
- We can select one country on a map and have evolution of count victims compare to mean of Europe indeed we have a kind of threshold
- Also, one point we think about is maybe to select one country on a map and have a kind of country disease pie chart such as follow.

So now let's move to a second section to focus on dashboard plots exploration and choose the 2 to add on our project.

4.3 Representations and interactions : Dashboard (evolution & victim informations)

Here is a kind of summary of which request and how we want it to interact with our visualizations.



- 2 more Req 3 designs

Req3 sol 1 en multiplot (≠ countries)

Req 3 Sol 2

Req 3 sol 1 ~~avec~~

points plottés en
X : qd le pays a + de malades
que moy EU

Req 3 Sol 3

O qd ——— − malades

+ / − :

~~Design first~~:

Rqmt 1 : − carte EU → positioning plots on
it will hide / be difficult
to [illegible] time
+ carte EU → shows geo dynamics
of diseases

Rqt 2 + dynamics in time due to x scale
= time
+ onClick → comparaison
country / EU

Rqt 3 − takes time to read information
due to symbols entropy ++
+ all features of data on
one graph (for one country
though).

Figure 8 – Sum up on what ideas guided us to our visualization

After our discussions in previous section, we first think about a tree map (Figure8) with all diseases but as we add a ListBox to choose the disease we want to focus, we turn tree map to a pyramid barplot to focus more about victim's characteristics (age, gender). So we turn this first idea from Figure8 to Figure9.

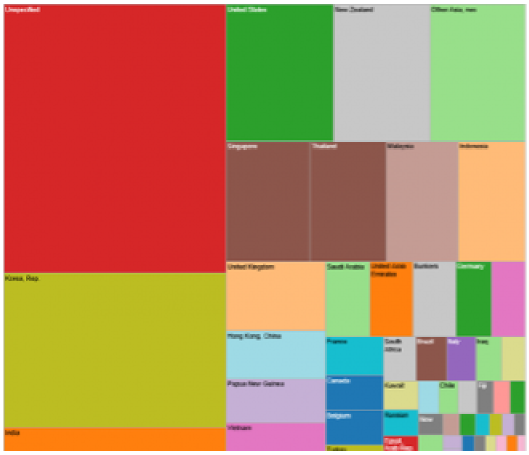


FIGURE 9 – Idea for the first plot

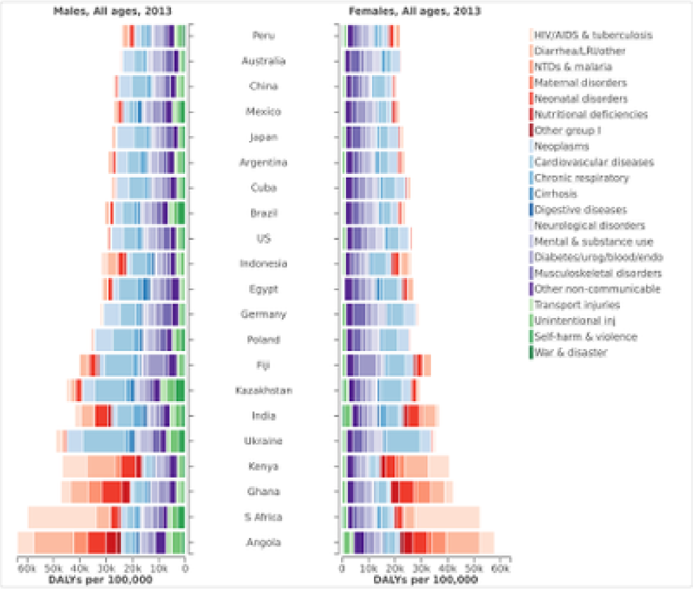


FIGURE 10 – Better idea fot the first plot

For the last graphic, we think that we missing a time representation to focus also on the evolution of one disease selected between all countries. So for that plot, we simply think about a time serie that represent each country disease evolution for 10 years range.

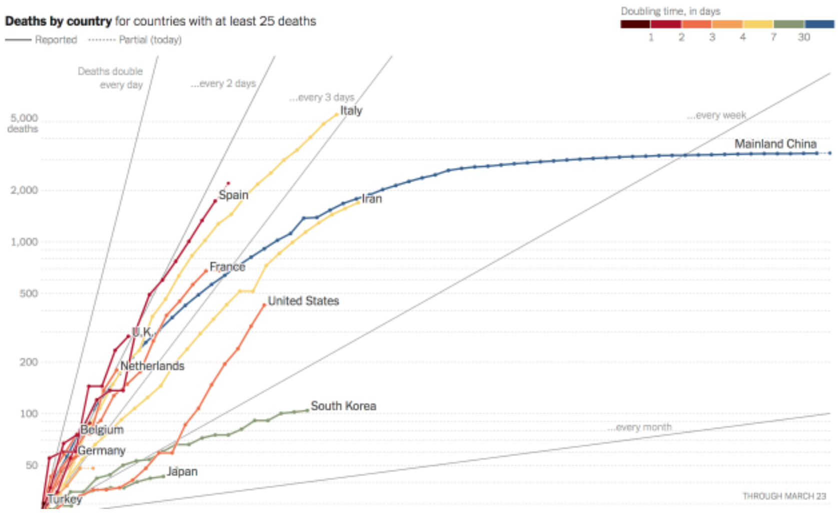


FIGURE 11 – Idea of graph for the second plot

To conclude, after all sketches and our drawings, we focus on a ListBox to choose one disease we want to focus. After that, we have our choropleth map depending on the disease selected and 2 plots. One focus on victim age gender repartition and another one focus on the evolution of the disease in 10 years range.

Now let's explain our code and make a presentation.

5 Final Presentation

To create our visualization we need a tool enabling us to link multiple views of the data to show the several information in our data set. As we experienced during previous labs, Altair is a good tool for combining several visualizations with an interactive dimension. **Altair** is a declarative statistical visualization library for Python, based on Vega and Vega-Lite.
After assessing our respective abilities with Altair, we agreed on a simplified visualization that answers the questions presented above. Our visualization is composed of three different views. For each view, data values are bind to a specific cause of death that we can choose via a drop-down list.
Our first view is a map with a color scale encoding that shows the total number of deaths for the European countries on the cause we selected. It allows us to have a quick look at the difference of number of deaths between countries : the colder the color is, the more number of death is caused by this cause in the country. To create this view, we used the "geoshape marks" tool for European countries map.

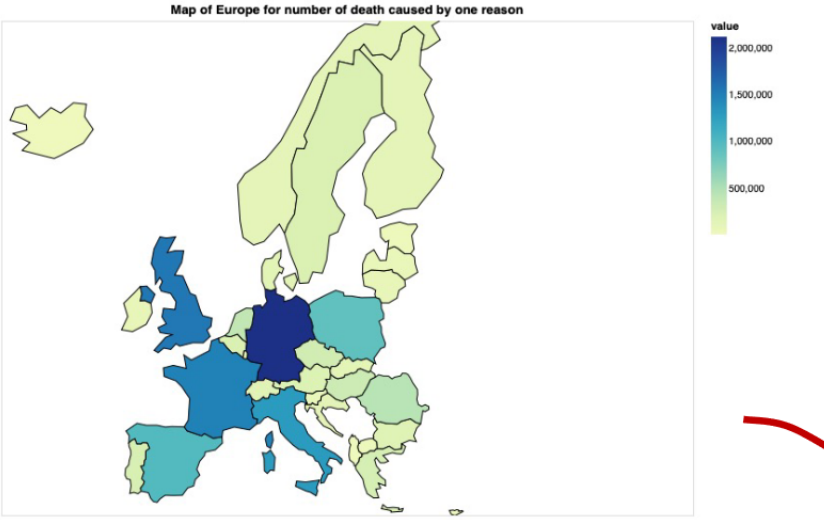


Figure 12 – First view

For this specific view, we needed to use an other data set and combine it with the first one. It contains country name and its corresponding **ISO 3166-1** numeric code for over 200 countries. It aims to visualise each country on the map for Altair.

Code	Country name
020	Andorra

The second view is a graph that shows the evolution of number of deaths due to one cause of death from 2001 to 2010. This graph helps us better understand the evolution of disease in each country and may improve the health plan for different countries. This view is also linked to the cause of death selected in the drop-down list.

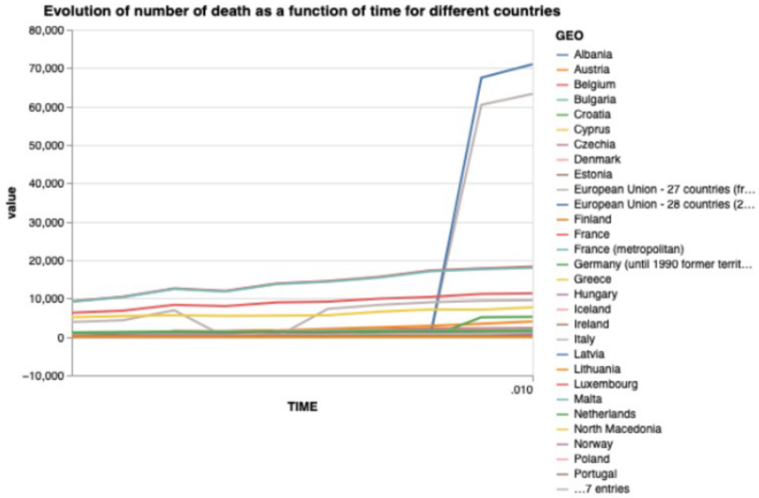


Figure 13 – Second view

The last view adds another aspect of our data. It allows us to compare between different genders. On the left we have the age distribution of number of death caused by one disease for male and on the right, the age distribution of number of death for female is shown. For example, we can see that the number of deaths caused by Alzheimer disease for female is higher the one for male, which can give the hint for hospitals and other institutions to pay more attention to female patient with Alzheimer.

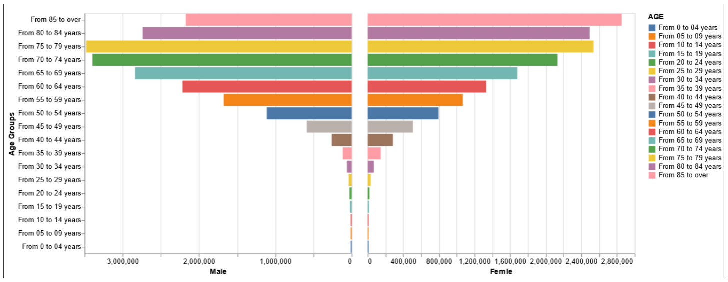


Figure 14 – Third view

6 Assets and limits

6.1 Mains qualities of our design

Our design does quite a few things well. This includes :

- Provide an intuitive visualization of disease repartition in the different European countries. Also go deeper in terms of discovery or comparison with the map, evolution of disease in each country, and for gender & age victims.
- Offer interactions permitting by selecting and focusing on one disease dynamically for all graphs.
- Provide an easy and guided navigation across the dashboard of our application.
- Offer a clear separation between each visualization and explore quite well each properties of our dataset. For example, map for country disease repartition, time serie for disease per country evolution, pyramid plot for victim characteristics.
- Offer serious and clever visualizations in our dashboard.

6.2 Limits of our design

Though quite exciting, our design may face a number of limits. For instance :

- One of our main limit is the dataset that is a little bit too old (stop in 2010), so our dashboard give a well and interesting exploration but not updated.
- Moreover, the time serie design struggles with sometimes strange scale and maybe we can have a better selection as TOP 5 most represented country for selected disease. Because 27 countries in the time serie can be difficult to separate and get insights.
- The information provided by the map is averaged for the whole population of one country, no matter if a region is closed to a border (this is inherent to the dataset). It will helped also to see how some epidemic / disease can cross Europe countries.

7 Conclusion

This project was a great opportunity for us to understand the process of creating a meaningful visualization from scratch, using the Altair python library.

We also had a glimpse of the several difficulties we may encounter in such projects, from the task of defining the question we want to help to answer to the kind of tools we can use.

By analyzing cause of death data, this project allowed us to put ourselves in the mindset of a health professional, and to question the usefulness of these data, as well as the different constructive questions that these users could ask themselves in order to have a real impact in their work.

Concerning the improvements we could make, it would be interesting to have an updated visualization of the causes of death to help our users to make good decisions.

Moreover, concerning the tool to be used for the analysis of these data, it would be preferable to create an interactive web page using “JavaScript” or “D3.js” for the visualizations. This will be more adapted to the different improvements listed in the Assets and limits section. This page will be able to ingest data in “csv” or other format from the WHO site.

On the data part, it would have been interesting to be able to visualize other variables, like the number of deaths according to a wealth cursor in order to check if the poorest populations are more at risk than the richest. This would have allowed us to see the effectiveness of public health policies in the different EU countries. It would also have been interesting to have the deaths on a regional perimeter, to have a finer visualization, and thus to be able to analyze more rigorously the impact of the diseases.

To conclude, we can make a link between our project and the visualizations during the COVID 19 pandemic. During this epidemic, we were able to realize the importance and usefulness of data visualization, which allowed the public authorities to really make important decisions for the population. However, as we have all seen, this powerful means of communication, that is data visualization, which is the fruit of the work of a data analyst or even a data scientist, must be interpreted by operational users, who will have sufficient business knowledge to make good interpretations.

8 Annex

Link to our Github project : `https://github.com/IGR204-Visualisation/Project-Milestone-IGR204`